Poisson regression case study: Household size in the Philippines

STA303 Winter 2021

This case study is drawn from the content in Chapter 4.4 of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). https://bookdown.org/roback/bookdown-BeyondMLR/.

Below is the code, with minor alterations, that accompanies Chapter 4.4.

```
library(tidyverse)
fHH1 <- read_csv(
  "https://raw.githubusercontent.com/proback/BeyondMLR/master/data/fHH1.csv") %>%
  select(-1)
```

Data Organization

```
glimpse(fHH1)
```

```
## Rows: 1,500
## Columns: 5
## $ location <chr> "CentralLuzon", "MetroManila", "DavaoRegion", "Visayas", "...
## $ age      <dbl> 65, 75, 54, 49, 74, 59, 54, 41, 50, 59, 72, 36, 42, 39, 65...
## $ total    <dbl> 0, 3, 4, 3, 3, 6, 5, 5, 6, 4, 2, 3, 7, 4, 5, 4, 2, 2, 2, 5...
## $ numLT5   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0...
## $ roof     <chr> "Predominantly Strong Material", "Predominantly Strong Mat...
```

Exploratory Data Analyses

```
mean(fHH1$total)
```

```
## [1] 3.684667
```

```
var(fHH1$total)
```

```
## [1] 5.534254
```

```
prop.table(table(fHH1$roof))
```

```
##
## Predominantly Light/Salvaged Material         Predominantly Strong Material
##                             0.1113333                             0.8886667
```

```
fHH1  %>% group_by(roof)  %>%
  summarise(mean=mean(total), sd=sd(total),
            var=var(total), n=n())
```

```
## # A tibble: 2 x 5
##   roof                                  mean    sd   var     n
## * <chr>                                <dbl> <dbl> <dbl> <int>
## 1 Predominantly Light/Salvaged Material  3.64  2.33  5.41   167
## 2 Predominantly Strong Material          3.69  2.36  5.55  1333
```

```
fHH1  %>% group_by(location)  %>%
  summarise(mean=mean(total), sd=sd(total),
            var=var(total), n=n())
```

```
## # A tibble: 5 x 5
##   location      mean    sd   var     n
## * <chr>        <dbl> <dbl> <dbl> <int>
## 1 CentralLuzon  3.40  2.04  4.15   224
## 2 DavaoRegion   3.39  2.17  4.72   187
## 3 IlocosRegion  3.59  2.32  5.40   191
## 4 MetroManila   3.71  2.21  4.86   297
## 5 Visayas       3.90  2.57  6.60   601
```

```
ggplot(fHH1, aes(total)) +
  geom_histogram(binwidth = .25, color = "black",
                 fill = "white") +
  xlab("Number in the house excluding head of household") +
  ylab("Count of households") +
  theme_minimal()
```

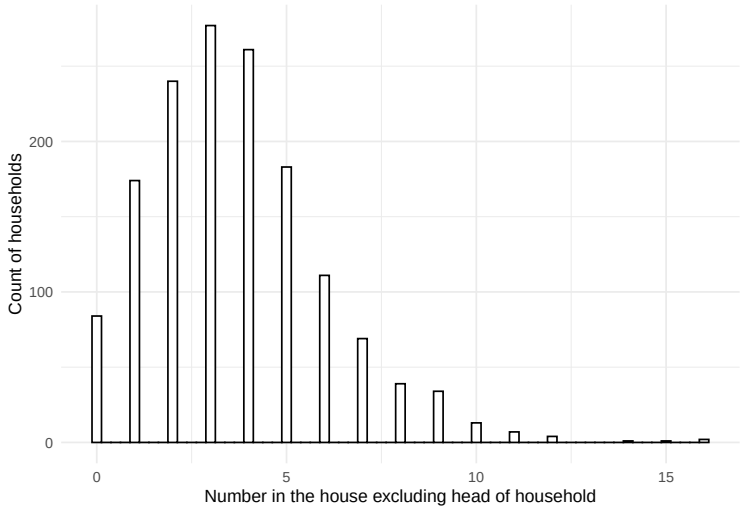

Figure 1: Distribution of household size in 5 Philippine regions.

```
cuts = cut(fHH1$age,
           breaks=c(15,20,25,30,35,40,45,50,55,60,65,70))
ageGrps <- data.frame(cuts,fHH1)
ggplot(data = ageGrps, aes(x = total)) +
  geom_histogram(binwidth = .25, color = "black",
                 fill = "white") +
  facet_wrap(cuts) +
  xlab("Household size") +
  theme_minimal()
```

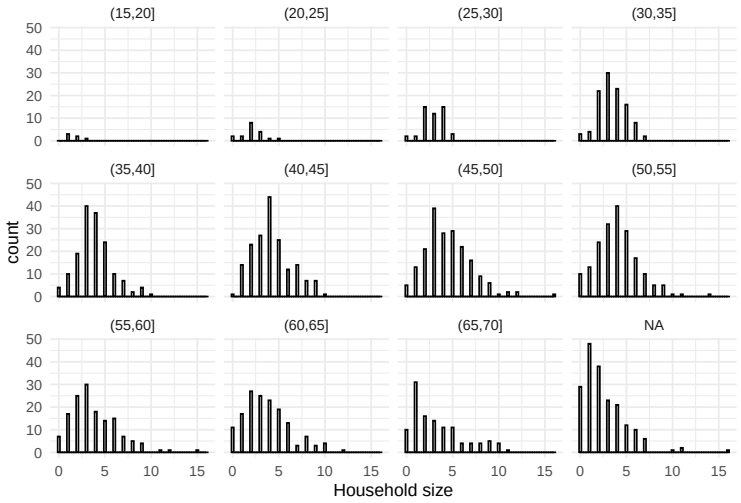

Figure 2: Distribution of household sizes by age group of the household head.

```
# Mean = Variance ?
ageGrps  %>%
  group_by(cuts)  %>%
  summarise(mnNum= mean(total),varNum=var(total),n=n()) %>%
knitr::kable(
      caption="Compare mean and variance of household size within each age group.",
      col.names = c("Age Groups", "Mean", "Variance", "n"))
```

Table 1: Compare mean and variance of household size within each age group.

Age Groups	Mean	Variance	n
(15,20]	1.666667	0.6666667	6
(20,25]	2.166667	1.5588235	18
(25,30]	2.918367	1.4098639	49
(30,35]	3.444444	2.1931464	108
(35,40]	3.841772	3.5735306	158
(40,45]	4.234286	4.4447947	175
(45,50]	4.489691	6.3962662	194
(50,55]	4.010638	5.2512231	188
(55,60]	3.806897	6.5318966	145
(60,65]	3.705882	6.1958204	153

Age Groups	Mean	Variance	n
(65,70]	3.339130	7.9980168	115
NA	2.549738	5.5435657	191

```r
## Checking linearity assumption: Empirical log of the means plot
sumStats <- fHH1 %>% group_by(age) %>%
  summarise(mntotal = mean(total),
            logmntotal = log(mntotal), n=n())
ggplot(sumStats, aes(x=age, y=logmntotal)) +
  geom_point()+
  geom_smooth(method = "loess", size = 1.5)+
  xlab("Age of head of the household") +
  ylab("Log of the empirical mean number in the house") +
  theme_minimal()
```

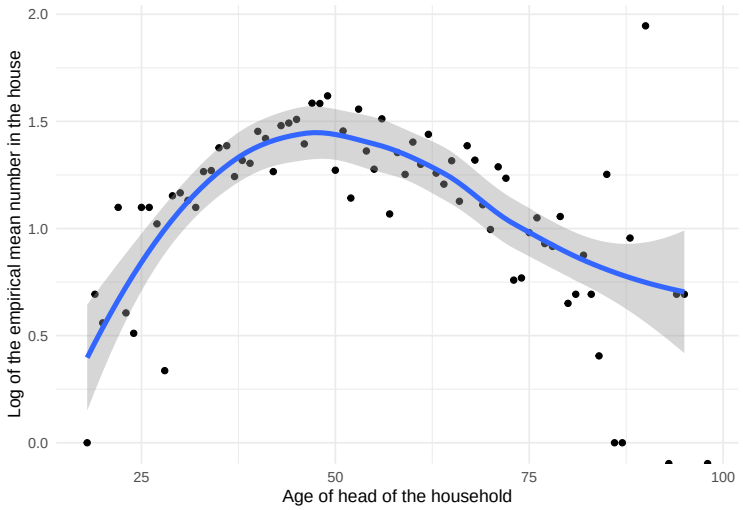

Figure 3: The log of the mean household sizes, besides the head of household, by age of the head of household, with loess smoother.

```r
sumStats2 <- fHH1 %>% group_by(age, location) %>%
  summarise(mntotal = mean(total),
            logmntotal = log(mntotal), n=n())
ggplot(sumStats2, aes(x=age, y=logmntotal, color=location,
                      linetype = location, shape = location)) +
  geom_point()+
  geom_smooth(method = "loess", se=FALSE)+
  xlab("Age of head of the household") +
  ylab("Log empirical mean household size") +
  theme_minimal()
```

```r
modela = glm(total ~ age, family = poisson, data = fHH1)
```

```r
coef(summary(modela))
```

```
##                 Estimate   Std. Error   z value      Pr(>|z|)
## (Intercept)  1.549942225 0.0502754106 30.829032 1.070156e-208
## age         -0.004705881 0.0009363388 -5.025832  5.012548e-07
```

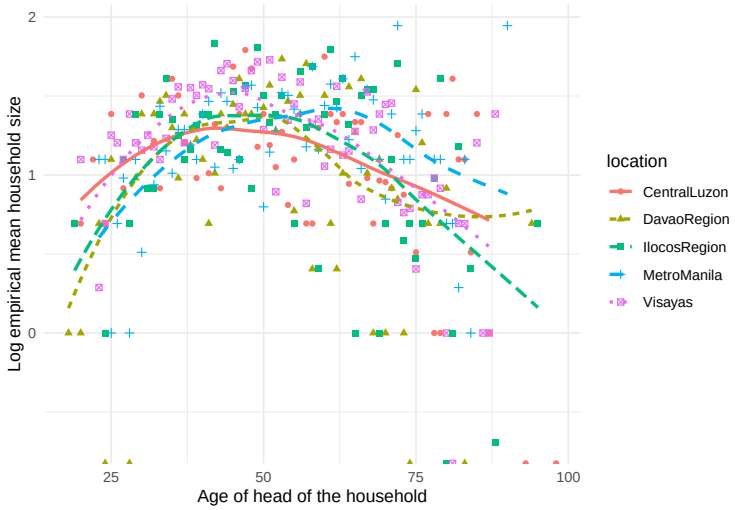

Figure 4: Empirical log of the mean household sizes vs. age of the head of household, with loess smoother by region.

```
cat(" Residual deviance = ", summary(modela)$deviance, " on ",
    summary(modela)$df.residual, "df", "\n",
    "Dispersion parameter = ", summary(modela)$dispersion)
```

```
##  Residual deviance =  2337.089  on  1498 df
##  Dispersion parameter =  1
```

```
# Wald type CI by hand
beta1hat <- summary(modela)$coefficients[2,1]
beta1se <- summary(modela)$coefficients[2,2]
beta1hat - 1.96*beta1se   # lower bound
```

```
## [1] -0.006541105
```

```
beta1hat + 1.96*beta1se   # upper bound
```

```
## [1] -0.002870657
```

```
exp(beta1hat - 1.96*beta1se)
```

```
## [1] 0.9934802
```

```
exp(beta1hat + 1.96*beta1se)
```

```
## [1] 0.9971335
```

```
# CI for betas using profile likelihood
confint(modela)
```

```
##                    2.5 %       97.5 %
## (Intercept)  1.451170100  1.648249185
## age         -0.006543163 -0.002872717
```

```
exp(confint(modela))
```

```
##                 2.5 %    97.5 %
## (Intercept) 4.2681057 5.1978713
## age         0.9934782 0.9971314
```

```
# model0 is the null/reduced model
model0 <- glm(total ~ 1, family = poisson, data = fHH1)
drop_in_dev <- anova(model0, modela, test = "Chisq")
```

```
did_print <- data.frame(ResidDF=drop_in_dev$`Resid. Df`,
    ResidDev=drop_in_dev$`Resid. Dev`,
    Deviance=drop_in_dev$Deviance, Df=drop_in_dev$Df,
    pval=drop_in_dev$`Pr(>Chi)`)
row.names(did_print) <- row.names(drop_in_dev)
did_print
```

```
  ResidDF ResidDev Deviance Df         pval
1    1499 2362.488       NA NA           NA
2    1498 2337.089 25.39907  1 4.661424e-07
```

Second Order Model

```
fHH1 <- fHH1 %>% mutate(age2 = age*age)
modela2 = glm(total ~ age + age2, family = poisson,
              data = fHH1)
```

```
coef(summary(modela2))
```

```
##                  Estimate   Std. Error    z value     Pr(>|z|)
## (Intercept) -0.3325296333 1.788357e-01  -1.859414 6.296847e-02
## age          0.0708867627 6.890442e-03  10.287694 8.007069e-25
## age2        -0.0007083289 6.405674e-05 -11.057834 2.008898e-28
```

```
cat(" Residual deviance = ", summary(modela2)$deviance, " on ",
    summary(modela2)$df.residual, "df", "\n",
    "Dispersion parameter = ", summary(modela2)$dispersion)
```

```
##  Residual deviance =  2200.944  on  1497 df
##  Dispersion parameter =  1
```

```
drop_in_dev <- anova(modela, modela2, test = "Chisq")
```

```
did_print <- data.frame(ResidDF=drop_in_dev$`Resid. Df`,
    ResidDev=drop_in_dev$`Resid. Dev`,
    Deviance=drop_in_dev$Deviance, Df=drop_in_dev$Df,
    pval=drop_in_dev$`Pr(>Chi)`)
row.names(did_print) <- row.names(drop_in_dev)
did_print
```

```
  ResidDF ResidDev Deviance Df         pval
1    1498 2337.089       NA NA           NA
2    1497 2200.944 136.1454  1 1.854452e-31
```

```
# Finding the age where the number in the house is a maximum
coefa2 = modela2$coefficients[3]
coefa = modela2$coefficients[2]
coefi = modela2$coefficients[2]
estLogNumHouse.f <- function(age){
  return(coefa2*(age)^2 + coefa*(age) + coefi)
}
optimize(estLogNumHouse.f, interval=c(20,70), maximum=TRUE)
```

```
## $maximum
## [1] 50.03803
##
## $objective
##     age2
## 1.844404
```

Adding a Covariate

```
modela2L = glm(total ~ age + age2 + location,
               family = poisson, data = fHH1)
```

```
coef(summary(modela2L))
```

```
##                           Estimate   Std. Error     z value     Pr(>|z|)
## (Intercept)          -0.3843337714 1.820919e-01  -2.1106581 3.480171e-02
## age                   0.0703628330 6.905067e-03  10.1900292 2.196983e-24
## age2                 -0.0007025856 6.420019e-05 -10.9436677 7.125764e-28
## locationDavaoRegion  -0.0193872310 5.378273e-02  -0.3604732 7.184933e-01
## locationIlocosRegion  0.0609819668 5.265981e-02   1.1580362 2.468493e-01
## locationMetroManila   0.0544800704 4.720116e-02   1.1542104 2.484139e-01
## locationVisayas       0.1121091959 4.174960e-02   2.6852758 7.246998e-03
```

```
cat(" Residual deviance = ", summary(modela2L)$deviance, " on ",
    summary(modela2L)$df.residual, "df", "\n",
    "Dispersion parameter = ", summary(modela2L)$dispersion)
```

```
##  Residual deviance =  2187.8  on  1493 df
##  Dispersion parameter =  1
```

```r
exp(modela2L$coefficients)
```

```
##          (Intercept)                  age                 age2
##            0.6809041            1.0728974            0.9992977
##  locationDavaoRegion locationIlocosRegion  locationMetroManila
##            0.9807995            1.0628797            1.0559914
##      locationVisayas
##            1.1186350
```

```r
drop_in_dev <- anova(modela2, modela2L, test = "Chisq")
```

```r
did_print <- data.frame(ResidDF=drop_in_dev$`Resid. Df`,
    ResidDev=drop_in_dev$`Resid. Dev`,
    Deviance=drop_in_dev$Deviance, Df=drop_in_dev$Df,
    pval=drop_in_dev$`Pr(>Chi)`)
row.names(did_print) <- row.names(drop_in_dev)
did_print
```

```
  ResidDF ResidDev Deviance Df       pval
1    1497 2200.944       NA NA         NA
2    1493 2187.800 13.14369  4 0.01059463
```

```r
modela4 <- glm(total ~ age + age2 + location + roof,
               family = poisson, data = fHH1)
summary(modela4)
```

```
Call:
glm(formula = total ~ age + age2 + location + roof, family = poisson,
    data = fHH1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9900  -0.9281  -0.1070   0.5912   5.0255

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -4.286e-01  1.865e-01  -2.298  0.02159 *
age                             7.040e-02  6.904e-03  10.198  < 2e-16 ***
age2                           -7.034e-04  6.419e-05 -10.958  < 2e-16 ***
locationDavaoRegion            -1.655e-02  5.384e-02  -0.307  0.75855
locationIlocosRegion            6.299e-02  5.269e-02   1.195  0.23194
locationMetroManila             5.322e-02  4.721e-02   1.127  0.25967
locationVisayas                 1.168e-01  4.196e-02   2.784  0.00537 **
roofPredominantly Strong Material  4.752e-02  4.359e-02   1.090  0.27564
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2362.5  on 1499  degrees of freedom
Residual deviance: 2186.6  on 1492  degrees of freedom
```

```
AIC: 6575.5

Number of Fisher Scoring iterations: 5
```

Residuals for Poisson Models (optional)

```
# Residual plot for the first order model
## Log scale
lfitteda = predict(modela) # log scale
lresida = resid(modela)  # linear model
lresid.df = data.frame(lfitteda,lresida)
ggplot(lresid.df,aes(x=lfitteda, y=lresida)) +
  geom_point(alpha = .25)+
  geom_smooth(method = "loess", size = 1.5, linetype = 2)+
  geom_line(y=0, size=1.5, col="red")+
  xlab("Fitted values") +
  ylab("Deviance Residuals") +
  theme_minimal()
```

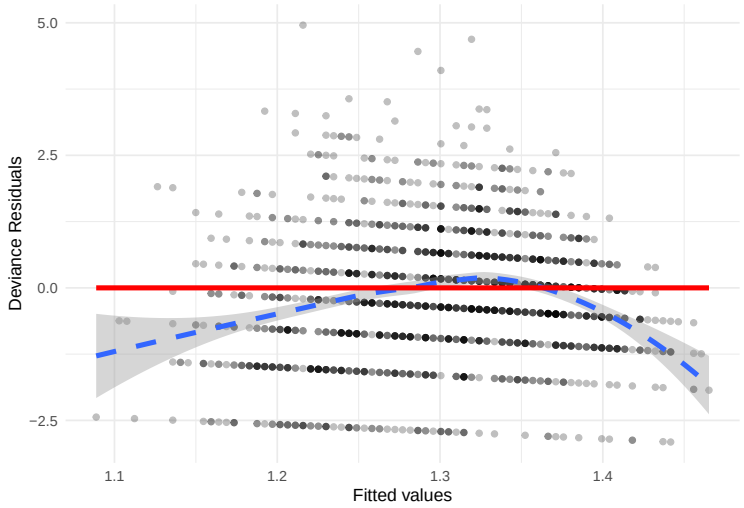

Figure 5: Residual plot for the Poisson model of household size by age of the household head.

Goodness-of-Fit

```
1-pchisq(modela2$deviance, modela2$df.residual)  # GOF test
```

```
[1] 0
```